ETL Proposal

Members: Alexander Mackenzie, Addisalem Abera and Dylan Grimm

Summary: We will be pulling LA Bike Sharing data (JSON) and active businesses (CSV) from Kaggle. The reason for this is because Dylan is going to LA at the end of the month and he wants to know if bike sharing locations are in good locations and close to different place he wants to go.

Links:

Bike Sharing:<https://bikeshare.metro.net/stations/json/>

Active Businesses:<https://www.kaggle.com/cityofLA/los-angeles-listing-of-businesses>

ETL Final

Extraction: We pulled LA Bike Sharing data (JSON) from their website and active businesses (CSV) from Kaggle. For links please reference above.

Transform: We cleaned up the CSV data so that it could work due to memory constraints. The raw CSV dataset encompassed 500,000 records and caused both encoding and memory errors. We trimmed the CSV data down to 15,000 records. We then condensed our data into a dataframe down to only having business name and addresses. We used pd.read\_csv, the copy function and renamed the columns since they spaces in between some column titles. With the JSON file we needed to open the file as an object using “with open” and json.load. We did this so we could iterate through the 2 different data frames that had a series of lists. This allowed us to pull the names of the different bike racks and the addresses associated with them. We appended the information into separate lists by tagging certain keys for the necessary information we wanted. We also checked the length of each series to make sure they were the same; we then referenced them to create our data frame.

Load: We loaded the non-relational data into MYSQL by using a create\_engine function to connect to MYSQL and then used our dataframe name with to\_sql function. This allowed to put the names of active businesses and bike racks with there associated addresses into our created tables so later analysis can be easily done to see if bike rack locations are in convenient areas and close to businesses Dylan is interested in visiting in LA. We created 2 tables (bike and businesses) that included an auto increment ID, with the names and addresses for the business and bike racks.